2021

Next-generation Protein Sequencing (NGPS) For Determining Complete Sequences for Unknown Proteins and Antibodies

Alexis S. Howard
University of Central Florida

NEXT-GENERATION PROTEIN SEQUENCING (NGPS) FOR DETERMINING COMPLETE SEQUENCES FOR UNKNOWN PROTEINS AND ANTIBODIES

by

ALEXIS S. HOWARD

A thesis submitted in partial fulfillment of the requirements
for the Honors in the Major Program in Biochemistry
in the College of Medicine
with the Burnett School of Biomedical Sciences
and in the Burnett Honors College
at the University of Central Florida
Orlando, Florida

Summer Term, 2021

Thesis Chair: Kersten Schroeder, Ph.D.

Running Title: Next-Generation Protein Sequencing

Keywords: Next Generation Protein Sequencing, antibody, sequence, proteins, polypeptide, cancer treatment, amino acids, proteomics, post-translational modification (PTM), diabetes, beta-cell dysfunction, diabetes mellitus (DM), type 2 diabetes (T2D), protein engineering, biotherapeutics

Table of Contents

Abstract

Next-Generation protein sequencing (NGPS) creates newfound ways of fully identifying every protein species in a single biological organism. It is an effort to use technology to determine proteomic data. The purpose of this research project is to use the current technology to sequence proteins and potentially find treatments for some diseases that are common today. Through NGPS, scientists can identify low abundant proteins including those that go through post-translational modifications (PTM) [1]. NGPS will allow us to fully determine protein sequences from protein samples using mass spectrometry with the ultimate goal of being able to determine the primary sequence of the protein in the given sample [1]. Antibodies are a specific class of proteins that aid our bodies in the immune response. Due to their variability in the complementary-determining region (CDR), NGPS will be used to determine the heavy and light chain sequences [2]. The goal of this technology is to fully determine the primary sequence of a protein in a given sample. The randomness of an antibody's variable (V), diversity (D), and joining (J) genes (VDJ recombination) makes each protein unique. VDJ recombination refers to the process of T cells and B cells randomly assembling different gene segments. This process allows the antibody to make specific receptors that can recognize different molecules presented on the surface of antigens. Proteases are enzymes that break down proteins and peptides. By using different proteases with varying cutting rules, we can digest the antibody and run it through high mass accuracy determining instrument [1]. NGPS allows us to utilize mass spectrometry technology to measure proteins or polypeptides. Because of these measurements, post-translation modifications, including glycosylation, can be detected, unlike in DNA sequencing technology. Protein sequencing has the opportunity to play a major role in the fight against the

COVID-19 outbreak and serve as curative measures for the treatment and Type 2 Diabetes [3]. Proteomics can serve as the basis of vaccine development as well as monitoring treatment. Utilizing techniques such as mass spectrometry could reveal the structure of the virus and ultimately allow for engineered tissues to produce the protein in large amounts in a lab setting. Currently, many companies are utilizing highly sensitized technology to carry out the goals of NGPS. The Oxford Nanopore is a company that uses technology to develop and explore more ways to undergo protein analysis. The methods used by this company involve using protein nanopores to mutate residues in pores to determine the overall sequence. The company utilizes modified aptamers that are drawn to the nanopore current. These aptamers can bind with some, but not all pores, allowing for the differentiation between target and non-target proteins. Nicoya Life Sciences is another company that uses Open Surface Plasmon Resonance (SPR) to detect molecular interactions. SPR uses an analyte (a mobile molecule) to bind to a ligand and observe changes in the refractive index. SPR allows researchers the ability to characterize the binding kinetics and affinities of monoclonal antibodies. SPR is an extremely promising technique to sequence proteins due to its flexibility in being able to work with a variety of molecules including lipids, nucleic acids, cells, viruses, nanoparticles, proteins, antibodies, carbohydrates, and more. The original goal behind NGPS was to establish a method to sequence proteins to aid in the early detection of common diseases such as Type 2 Diabetes. After significant research, it is now known that NGPS can be done in a variety of ways to accomplish a common goal—sequencing proteins and understanding how amino acids affect the human body. In the case of diseased states, NGPS can help researchers find ways to diagnose, treat, and cure diseases early on. Focusing on antibodies allows us to manipulate the body's immune response to rid the host

of pathogens. NGPS, however, is advancing at a much slower rate than anticipated by companies due to its many limitations including not being able to sequence large peptides, difficulties in material and composition of the sample, and needing to label small peptides to begin degradation. Ideally, finding a way to combine the high accuracy and specificity of certain techniques, the ability to detect low abundant proteins in others, and the flexibility of Open SPR would allow researchers and companies to create the standard for NGPS. Creating effective antibodies is precisely why NGPS has such great potential today. Ultimately, I found that as a standalone, Open SPR is the most effective method. However, as the research shows, there are limitations with each method, including Open SPR. The conclusion shows that it is necessary to find a combination of these techniques and create an accurate method, potentially using different technologies, to create the most effective way to sequence proteins.

Introduction

Immunoglobulins (Igs) are proteins produced by the immune system that neutralize pathogens that invade the human body. Igs recognize an antigen by way of the fragment antigen-binding (Fab) variable region. Because antibodies are a remnant of the immune response that ultimately aids in the body's response for similar future ailments, understanding what antibodies are in our systems will allow us to know what has entered our bodies. Next Generation Protein Sequencing will determine the primary sequence of proteins employing mass spectrometry. NGPS is a renewed approach towards identifying and quantifying every single protein species in complex biological organisms [3]. The uses of this technology range from biological studies to clinical applications to be able to better understand the causes of various diseases including Alzheimer's and Parkinson's to raise earlier detection and provide more effective treatments for those suffering. What ultimately makes this study so difficult, in comparison to Next Generation DNA Sequencing, is that there are only five different bases that can interact between DNA and RNA whereas proteomics comprises 20 different amino acids that go through post-translational modifications. One of the most important pieces of information researchers need in antibody-drug research and development is a deep understanding of the protein. Oftentimes, researchers conduct sequence analysis using the advancement of new mass spectrometry technologies. The likelihood of an individual being predisposed to the few diseases above increases drastically with family history and environmental stimuli. One of the largest things that play a role in our bodies as humans are single nucleotide polymorphisms (SNPs). SNPs are essentially what allows us to differ from each of our parents on a genetic level. Genetically speaking, we obtain 50% of our DNA from our mother and the other 50% from our father. These SNPs are the only differences

that can account for why we aren't 100% similar to both of our parents. Because SNPs occur at such a high rate in comparison to mutations (less than 1% of the population contains mutations), SNPs can be traced back to some of these diseases and disorders that the F1 generation may experience differently from the parental generation. The ability to sequence our proteins and study SNPs has the potential for researchers to find and edit these changes. A multitude of changes can affect the type of protein produced and consequently, how the body responds to these changes, and these changes are named mutations. Mutations are an alteration of the base order which may result in variable effects. The types of mutations that can occur include a silent mutation, in which there is a change in the codon base sequences that ultimately does not change the protein it codes for. Another is a missense mutation which changes a single base in the codon sequence which can alter the type of protein produced. Oftentimes, when a different protein is produced but falls under the same family as the intended protein, very little change is assumed. A nonsense mutation may occur due to a deletion or insertion of a nucleotide that can cause a truncation of the genetic code, leading to a stop codon being prematurely added to the sequence. Lastly, a frameshift is a type of mutation that may cause a stop codon to be added or even a different protein to be added due to a different reading frame. As discussed in-depth, later on, there are five (5) categories that constitute where the 20 amino acids fall due to the variable side chain that differs among each one. There are nonpolar, polar, basic or positive, acidic or negative, and aromatic amino acids. Each of these categories possesses its character that may lead to phenotypic changes and disease of the organism experiencing the alteration of the genome. DNA and RNA go hand-in-hand with the polypeptide sequence that is synthesized in organisms. As coined by the Central Dogma of Molecular Biology, DNA is first transcribed into

RNA, which is translated into proteins. These proteins are what essentially govern the body and drive all of its biochemical processes. Errors in these processes may lead to inactivation and can result in or lead an organism to become prone to specific diseases such as non-insulin-dependent diabetes mellitus (NIDDM), cancer, or most recently, the 2019 novel coronavirus (COVID-19).

Methodology

Researchers have found varying ways to determine the best method to study these protein sequences. Affinity assays are vital to clinical and research biology because they generate a great deal of critical information. Affinity assays are a fast, cheap way to determine proteins and can be very sensitive. Similar to an antibody, these assays can only capture a single molecule at a time. Assays are classified as qualitative data, which makes it hard to analyze and measure multiple protein species in the same assay. Mass spectrometry (MS) is a more reliable way to determine protein sequences as it is precise and can determine the abundance of a single sequence in a complex biological organism. MS has its limits, for example, it requires nearly 1 million molecules from a homogenous sample to fill its chambers, which jeopardizes samples with lower concentrations. It has been found that to identify the antigens on a tumor sample, an estimated 100 million- 1 billion cells were needed, which is well over the amount available from a biopsy [1]. RNA sequencing provides accurate quantification of molecules within a sample however, it has been proven to deliver an indirect measurement of protein expression. MS has shown to be the most popular and sensitive method for protein sequencing. With access to a cell line, NGPS, utilizing MS, can detect unexpected variants in antibodies. RNA and DNA sequencing technology cannot detect protein changes, and, in some cases, this technology may not work on most antibodies due to the hyper-variable CDR region on the antibody [1]. The introduction of REmAB technology will deliver accurate heavy and light chain full-length sequences but fall short due to the fact that Leucine and Isoleucine have the same mass. This, however, in combination with W-ion Isoleucine/Leucine Determination (WILD™) technology can account for this as it is a technology that is designed to distinguish two isomeric amino acids

using MS. WILD™ technology utilizes the most advanced mass spectrometer to determine Leucine and Isoleucine more accurately based on the w-ion observed in the experiments [3]. WILD™ can determine every Leucine and Isoleucine position in the variable region of heavy and light chain sequences since they vary in structure. Intact mass analysis is another form of measurement that determines the intact antibody-protein molecular weight. This technique can measure the protein in its native form, or it can measure the heavy and light chains separately after reduction. This method can provide extra information to identify and confirm the primary sequence of the antibody from a known primary sequence. Variations such as Leucine/Isoleucine mutation cannot be determined through this method and WILD™ technology can determine those changes in sequence. Intact mass analysis can derive ratios of expressed glycoforms that confer biological activity or immunogenicity [3]. In performing an intact mass analysis, steps include performing an LC-MS on purified antibody protein or its fragments to generate full MS spectrum, followed by deconvolution of the high charged full MS spectrum, and data analysis and interpretation [3]. Peptide mapping is another method that analyzes peptides made from the digestion of a protein by MS. This is a technique not used as often since it requires a reference sequence to compare and contrast. However, it is most useful in confirming the sequence of a target protein against a known sequence to discover point mutations. Because peptide mapping utilizes more information than intact mass measurements, it is seen as a more reliable method. At the start of my research, I began looking at keywords that included peptide mapping, protein sequencing, and antibodies. The original goal was to understand how these antibodies, or rather how sequencing these antibodies, could aid researchers and developers in finding more effective treatment and vaccinations for current

illnesses. Upon searching protein sequencing, and the methods that accompanied it, I found that many other illnesses can be treated and potentially cured with understanding and manipulating peptides. Many papers discussed protein therapeutics for the treatment and preventative measures in those that carried specific genes for cancer. I found that protein therapeutics has dominated the medical industry for the past few years. I investigated the different technologies and techniques that are currently being pursued for identification (as seen in T2D), treatment (as seen in vaccinations), and curative or preventative measures (as seen in Cancer treatment). After finding the different techniques being used, including but not limited to the aforementioned technology such as affinity assays, mass spectrometry, and WILD™ technology, the main question that arose was, “Are there more methods?” Through further research, I found Surface Plasmon Resonance (SPR), the Oxford Nanopore, Recognition Tunneling, Subnanopore Sensoring, and Image-Based ClpXP Digestion. I began to look into the differences between Image-Based ClpXP Digestion and Edman Degradation, a known technique that has been widely used in the past. After understanding the basis of the different techniques, the question to answer became "Which technique is going to provide the most accurate results most effectively?" Looking through more databases and reading more articles, I concluded that no one technique was the 'best.' Each technique and technology had its highlights that also came with its fair share of limitations. Moving forward, I investigated how NGPS could be applied to diseases that exist today and delivered my recommendations on which combination of techniques could provide the best answer to my question.

Applications

Diabetes Mellitus

Type 2 Diabetes (T2D) Mellitus is a common disease affecting more than 400 million people globally and 34.2 in the U.S alone. T2D Mellitus is the seventh leading cause of death in the United States and the number 1 cause of kidney failure, lower-limb amputations, and adult blindness. Sequencing technologies such as whole-exome and whole-genome sequencing applications have provided information into the molecular bases of T2D Mellitus [4]. The implementation of next-generation sequencing will shed light on new knowledge about T2D Mellitus genetics in diagnosis, prevention, and treatment of the disease [4]. The disease is characterized by systemic insulin resistance and beta-cell dysfunction. As the combination of the two exacerbates, hyperglycemia amplifies leading to T2D [5]. The etiology of T2D is known to have a component largely associated with genetics confirmed by family and twin-based studies. Studies have shown that the risk of developing the disease increase from 40% to 70% when one or both parents have the disease, respectively. The rate increases to as high as 76% in monozygotic twins. Early identification of high-risk individuals allows for the delay and potential prevention of T2D onset through lifestyle interventions and has consequently reduced costs of healthcare for those individuals [4]. Published studies have placed emphasis on the identification of T2D susceptibility loci from next-generation sequencing (NGS) data. In a Danish study in which 2657 Europeans with and without T2D were used as subjects, results showed that the variants associated with T2D were overwhelmingly common and most located

within regions identified by Genome-wide association studies (GWAS) in an effort to investigate the hypothesis of "missing heritability" [4]. The NGS technology determined that a coding variant reached genome-wide importance that was shown to be common in East Asian ancestry (PAX4). The PAX4 gene encodes a transcription factor involved in islet differentiation and function and NGS technology has shown some PAX4 variants have been associated with early-onset monogenic diabetes [4]. Overall, there are several paths in which NGS may be used to assist in the identification of causal genes and variants for T2D, but limitations are still evident. Despite the decreasing costs of NGS-based approaches, the most evident limitation is the difficulty of obtaining enough observations to make a valid statistical inference [4].

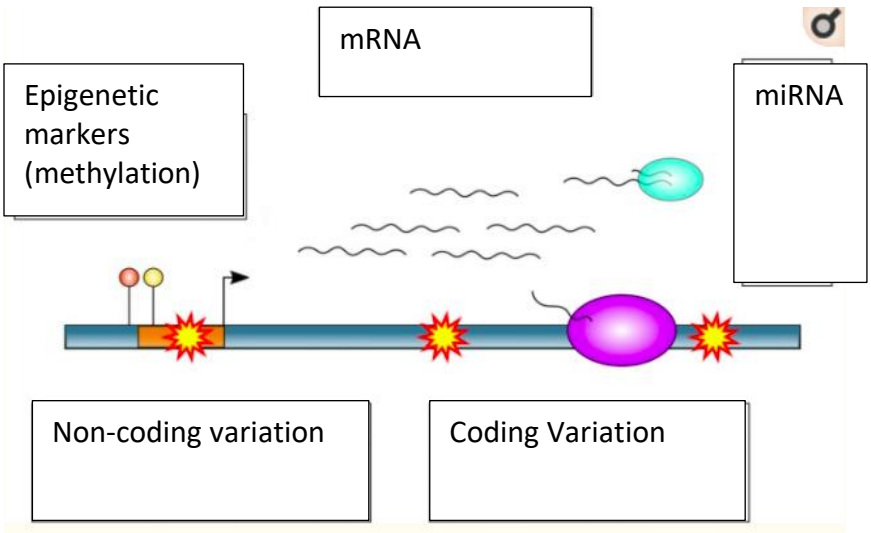


Figure 1. T2D Loci

Adapted from [4].

Cancer Treatment

Protein therapeutics for cancer treatment have dominated the medical industry in the last few years. Currently, most approaches for the development of protein-based cancer treatment utilize features of the tumor and tumor-associated cells to attack the overexpressed cells [6]. Protein therapeutics can actively target cancer cells by promoting destruction by binding to cell surface receptors and others marked with associated tumor cells in comparison to healthy tissue. Upon binding to the ligands on the overexpressed tumor cells, the proteins can deliver apoptotic signals to induce cellular death. Cellular death is induced by blocking cellular signals important for survival. Many prominent therapeutics act by localizing cytotoxic agents directly to tumor cells [6]. Another method to induce cell death is by targeting the vasculature and stroma of the tumor cells. The stroma of these cancer cells contains structures required for cell growth and development such as myofibrils, blood and lymphatic vessels, inflammatory cells, etc. Without this undisturbed microenvironment, tumor growth, progression, and metastasis are unlikely to occur. By targeting the stroma and vasculature of tumor-associated cells, proteomics can potentially prevent tumor survival and proliferation. Other protein therapeutics can function through opsonization by eliciting immune-mediated killing via complement, induction of phagocytosis, or changes in T-cell function [6].

SARS-CoV-2

Next Generation technology focuses primarily on sequential information. The outbreak of the SARS-CoV-2, the severe acute respiratory virus widely referred to as coronavirus, is predicted to have 133 vaccines currently underway of development. The simple advantage to developing a next-generation vaccine is that it is based on sequencing alone. Current vaccinations depend mainly on the time it takes to culture the organism causing the virus. In the case of the SARS-CoV-2 genome, large quantities of the virus would need to be grown for vaccination to occur [6]. Next-Generation Protein Sequencing can protect from infection or disease by manipulation of the known protein sequence of the associated spike protein. As new variants to the original virus arise, this next-generation technology will have the means to manipulate the genome. The current SARS-CoV-2 is 40% genetically similar to the SARS-CoV that was seen in 2002. See Figure 2 adapted from [6].

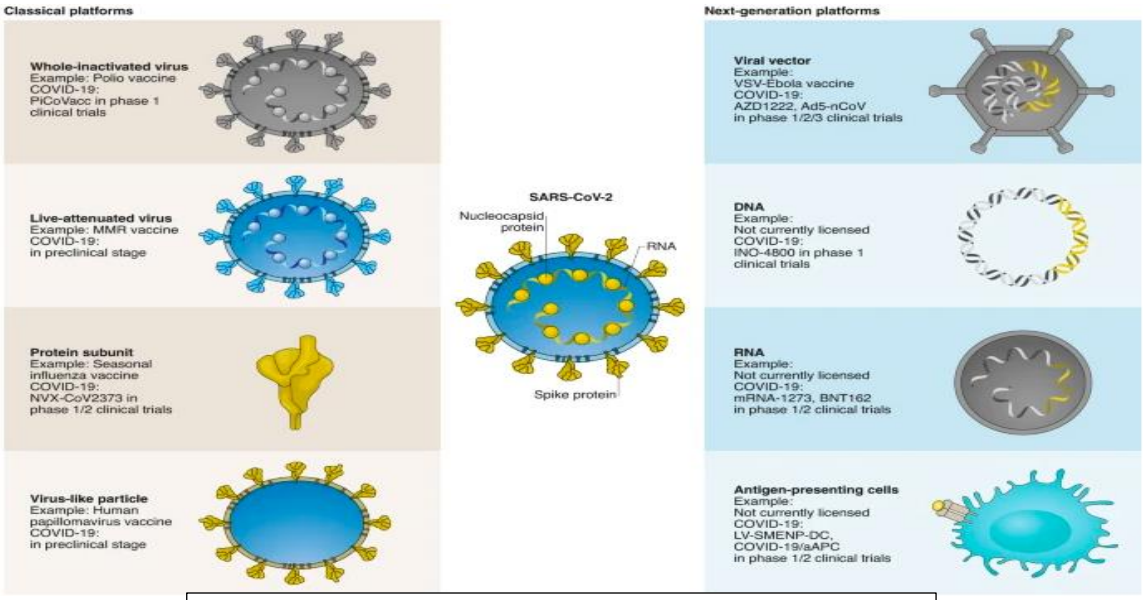


Figure 2. Classical Vaccine Routes for SARS-CoV-2 Spike protein

Discussion

Recommendations

Next Generation Protein Sequencing is a technology that although can be the main source of breakthrough for many scientific medical studies today, remains fairly new and untouched. Next Generation DNA Sequencing has been used for the past few years to sequence thousands of genes and even entire genomes in a relatively short time frame. Thus, this technology has paved the way for detecting mutations and variants in disease. As a result, it has been used in diagnosis, prognosis, and follow-up care in many patients to date [7], but NextGen DNA Sequencing has its shortcomings—and that's where NGPS comes in. With the funding and access to technology, researchers would have the ability to identify every protein species and its sequence in a single biological organism. This would give rise to understanding many incurable diseases and how those diseased states affect proteomics. Ultimately, by determining which antibodies are present in a given organism, could give answers as to which pathogens we have previously encountered and the likelihood of its ability to spread would come to light. This technology could have been responsible for potentially reducing the effects of the current COVID-19 pandemic to only an epidemic, decreasing its' spread, allowing us to get ahead of its effects. Because this technology is De Novo, it doesn't rely on a database, making it able to sequence any protein of any isotope. Owing to the little research done on NGPS, much information on the technology is unbeknownst to us. In other words, NGPS can be the next biggest thing in medicine in the coming years.

Structure

To break it down to a science, let's first look at the importance of proteins. Amino acids are the building blocks of proteins and the structure of which includes an organic central carbon atom connected to an amino (-NH2) group known as the N-terminus, a carboxyl group known as the C-terminus, a Hydrogen atom, and a variable region known as a side chain. These side chains vary and based on the structure of the side chain, it is what gives one of the 20 (and 21, if you count selenocysteine) amino acids its' name.

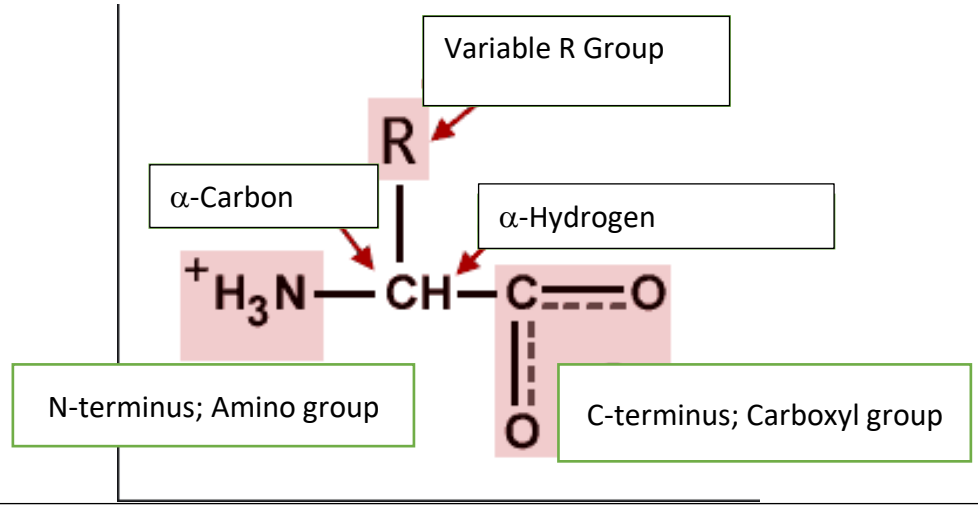


Figure 3. Base structure of amino acids. The R-group refers to the unique side chain, which gives the amino acid its properties such as polar or nonpolar (hydrophobic), acidic, or charged.

The type of side-chain each amino acid contains helps determine its properties and thus, function. One of the taglines in the broad subject of science is "Specific shape, specific function" for this reason. For example, an amphipathic protein with both a polar and nonpolar

(hydrophilic and hydrophobic, respectively) region would fold in a way such that the nonpolar region remains on the inside of the protein and is never in contact with water. These amino acids bond together to form proteins, which have their own specific structures. The four structures of proteins include the primary, secondary, tertiary, and quaternary structures. The primary structure is specifically important in this study because due to being known for its peptide bonds, the 1 deg structure ultimately determines the function of the protein. There are several conditions which can reverse the folding of proteins including organic solvents, chaotropic agents, pH, and temperature however, the peptide bonds in the primary structure cannot be denatured. This means, essentially, that the main function cannot be lost. The secondary structure of proteins is important for its alpha helices (stabilized by H-bonds) and beta-pleated sheets. This can be noted in the cause of Mad Cow's disease. Mad Cow's disease is a disorder due to evidence of prions, or misfolded proteins, in which the structure is altered from alpha-helices to beta-pleated sheets. The tertiary structure is known for its interactions of the unique side chains previously discussed and lastly, the quaternary structure is responsible for assembled units.

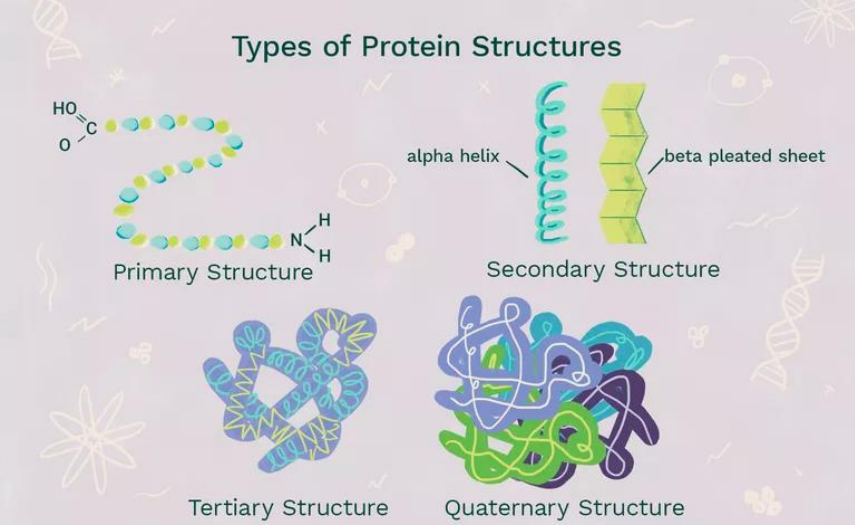


Figure 4. Structures of proteins. Note the clear distinction of the C-terminus vs the N-terminus on the primary structure.

Function

Now that we understand the structure, what about the function? Proteins, as we now know, adopt a specific 3D conformation which is known as its native fold. This native fold is associated with a small cost in conformational energy but allows for the protein to adopt a shape that gives it the ability to make a large number of favorable interactions within the protein. Because of the wide variability of proteins, they possess many vital functions in the human body. From hair to the immune response, proteins can make us extremely healthy and fit to lead a long life. However, this does not come without a price. Proteins also can make us just as sick and unwell. Take a look at one of the most (in my humble opinion) proteins in the body: hemoglobin. At a young age, they teach us that hemoglobin (Hb) is what makes our red blood

cells red, but there is so much more to Hb than simply that. Containing a heme group, hemoglobin contains a protoporphyrin ring that interacts with four out of the six coordination bonds of iron (forming the heme). Hb is a tetramer that is also responsible for binding four Oxygen (O2) molecules at a time. These O2 molecules are required for the basic functions of the body including but not limited to neurological function, musculoskeletal movement, renal function, cardiac function, hepatic enzymatic function, and much more. Due to the high affinity of O2, Hb is an extremely important protein in the body but comes with a series of health complications when not at maximum function. Anemia is one of the more common disorders or Hb, more specifically, Sickle-cell anemia. Sickle-cell anemia is a disorder in which the beta-globin is mutated where valine (V) is substituted for glutamic acid (E). This causes an irregular shape of the erythrocytes and thus, each RBC only carries half the normal Hb. This can lead to a lack of O2 being circulated in the blood throughout the body. Thalassemia is another disorder occurring from an abnormal amount of hemoglobin due to deletion. Another type of protein that the body uses are antibodies (Ab). Antibodies are the product of the body's immune response to invading pathogens. As discussed earlier, antibodies play a crucial role in ridding the body of these disease-causing organisms in adaptive immunity as well as in vaccination. Each of the protein structures pictured above has different characteristics. The primary structure, known for its peptide bonds, is a series of amino acids added to the C-terminus of the previous amino acid. The primary sequence is one of the most important in determining which protein will be produced. As displayed by Christian Anfinsen in 1973, he executed the Ribonuclease Experiment with Ribonuclease A, which showed that denaturing the protein in a non-harsh way would allow the protein to refold, following its' original primary sequence. The experiment

concluded that the primary structure of the amino acid sequence contains all the necessary information for the renaturation of the protein under optimal standards. This did not stray away from the idea that many proteins require molecular chaperones to prevent improper folding. For example, protein disulfide isomerase (PDI) forms and breaks disulfide bonds in the endoplasmic reticulum (ER) to assist in folding. The secondary structures are known for their α-helices and β-pleated sheets creating coiled coils. It is important to note that proline and glycine are α-helix breakers. Proline tends to cause turns in the helix as it is not able to form Hydrogen bonds (H-bonds) whereas Glycine is too small and flexible due to Hydrogen as an R group. These structures are not the only secondary structures, but they are the most stable and therefore, most common. They play major roles in the human body. For example, keratin, the main protein found in human hair, is a right-handed α-helix. The palm region of DNA Polymerase, the main enzyme involved in DNA synthesis, is a β-pleated sheet. β-pleated sheets come in two main conformations which include the parallel and antiparallel configurations, with the antiparallel being more stable than the former. As labeled in Figure 3, the R group, or side chain, of an amino acid is variable. These side chains play an important role in understanding the tertiary structure of a protein. There are five groups of amino acids in which the variable R group determines—hydrophobic, polar uncharged, acidic/negative, basic/positive, and aromatic. The tertiary structure is an interaction of amino acid side chains, and thus folds accordingly to the type of amino acids. The tertiary structure folds to maximize H-bonding and other weak interactions while utilizing the lowest Gibb's free energy. Levinthal's paradox states that it is nearly impossible for every protein to try every conformation to fold with the lowest energy. It would take years for this to occur, yet proteins can fold in seconds. The quaternary structure is

an interaction of interactions. It uses a variety of weak forces including van der Waals, Hydrogen bonding, and hydrophobic effects to arrange the protein subunits favorably. One of the strongest influences on protein folding is the hydrophobic effect, a spontaneous reaction with a $-\Delta G$. Due to this, proteins are likely to fold with nonpolar amino acids on the inside to minimize interaction with water. Without this, many diseased states can occur including but not limited to Sickle cell anemia, as previously discussed. Proteins contain a series of bundled secondary structures which account for a motif, a nonfunctional region of the molecule. Domains, functional units on the protein, are typically seen in tertiary structure. These domains are independent sites of a larger protein molecule that carry out varying functions on the same molecule. Oftentimes, catalytic sites are found in between domains on a protein.

Protein Sequencing using Mass Spectrometry

As gel electrophoresis is being phased out, we are now coming into new methods of quantifying the amino acid sequence in a protein. Gel electrophoresis is a method used to separate DNA fragments into loaded wells through charge. This technique was developed by filling the assay with gel and utilizing an electric current to pull the negatively charged DNA to the positively charged electrode. This process aided in the ability to distinguish DNA based on size and even charge. As we shy away from gel electrophoresis, new techniques such as mass spectrometry have allowed us to make advancements in analyzing data. Mass Spectrometry (MS)

can study the protein by chemically cross-linking fragments and analyzing the results. MS digests the protein, analyzes the fragments, repeats this process with different enzymes, and rebuilds the protein. This is cost-effective for an entire protein sequence by isolating and sequencing its DNA or mRNA and is also fairly fast. Because MS is a mass-to-charge (m/z) ratio of the molecules present in a sample, we can understand different characteristics of the amino acids for purposes of identifying them. These numbers are often utilized to calculate specificities of the molecules including molecular weight, protein interactions, post-translational modifications, and protein quantification. Through the usage of reverse-phase liquid chromatography (RP-LC) and tandem MS, small peptide fragments identified by MS are matched to protein databases for the determination of the amino acid sequence.

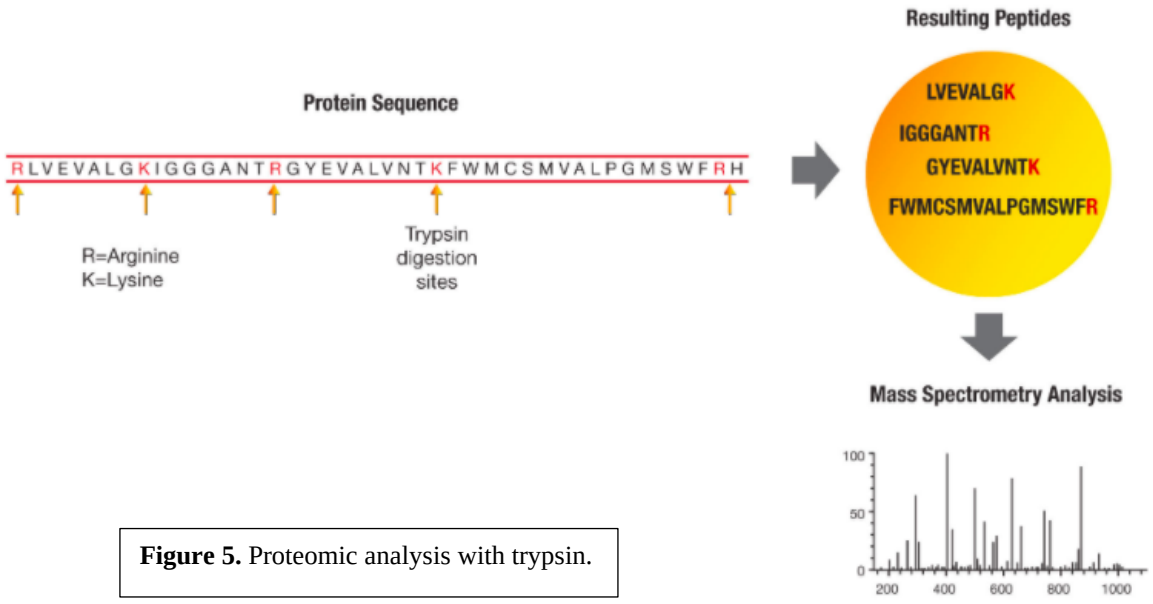


Figure 5. Proteomic analysis with trypsin.

Subnanogap sensoring utilizes electrophoresis along a constrained path. It is estimated from the number of steps in the trace with an average of 90-95% accuracy. Recognition tunneling is another severely underdeveloped approach. An advantage to it is having the ability to create a computational pattern recognition of current trace that has >99% accuracy. Image-based ClpXP digestion is a method that is fairly underdeveloped but provides an estimation of the chain length of bound protein with a rough efficiency of 70-80%. Its attachment of fluorescent dyes to cysteine (C) and lysine (K) provides for 95% efficiency in its side-chain assignment. Image-based Edman degradation provides for reactions of carboxylic acid groups (-COOH) with amide on the surface with approximately 95% efficiency.

Subnanocap Sensoring

Subnanoscale pores are a method in which a pair of electrodes establish an electric circuit on opposite sides of a microfluidic cell. As we quantify the total analytes move through the pore, the total number can allow for quantification of similar species. The sensitivity of these nanopores can be used to distinguish between these similar peptides through electric fingerprints. Because this technology primarily focuses on volume, even small amino acids could be distinguished. The main disadvantage of this technique is the length of time required to achieve this goal.

Recognition Tunneling

Recognition tunneling utilizes small particles passing through an energetically unfavorable region. Because of its high accuracy (>99%), recognition tunneling is extremely popular among de novo sequencing for purposes of identifying amino acids. One of its limitations, however, stems from its ability to only work for free amino acids.

Image-Based ClpXP Digestion

In accordance with the high frequencies of lysine (K), this technique can be applied to almost every protein sequence, allowing for great variability. The timeframe for this method would be relatively short, unlike in recognition tunneling, because the distinction in abundance can be made without prior separation. The main limitation with this technique would be the need to label the C-terminus with a small RNA A peptide to begin digestion. Edman degradation is a method of sequencing amino acids in a peptide. During Edman degradation, the N-terminal is labeled and cleaved from the peptide without disrupting the peptide bonds between the other amino acid residues. Edman degradation uses phenyl isothiocyanate as its reagent to aid in the process of purifying proteins by removing one amino acid residue at a time from the amino end (N-terminus) of a peptide chain. Edman degradation is a useful technique because it allows for

the end amino acid to be cleaved in less time without damaging the overall peptide. In Edman degradation, we are not able to fully sequence and cleave the amino acid residues at the N-terminus on large proteins. As a result, it is required that we cleave larger proteins into smaller ones and begin sequencing the smaller ones.

Oxford Nanopore

Oxford Nanopore is currently a company that utilizes technology to explore and develop more methods for protein analysis. It is a process by which protein nanopores are mutating residues in the pore to determine its sequence [9]. This technology can benefit researchers looking to explore and validate new proteins since it is designed to analyze sequencing with high specificity and high sensitivity [9]. Oxford Nanopore technology utilizes modified aptamers (oligonucleotides that bind with the target protein of choice) [9]. This is then drawn into the nanopore to disrupt the current. Once this occurs, the aptamer is ready to leave and interact with another fragment. Because each protein fragment does not bind to the nanopore, this method can differentiate between target and nontarget proteins [9].

Surface Plasmon Resonance

Surface Plasmon Resonance (SPR) is an optical technological technique that is used to detect molecular interactions. SPR utilizes the binding on an analyte, which is a mobile molecule, to a ligand to observe changes in the refractive index. When polarized light strikes a surface, the plasmons (electron-charged waves) decrease the intensity of reflected light to the mass on the surface. Many times, gold is used in SPR due to its useful properties of angle and wavelength. Gold is also inert (non-reactive) to solutions that are commonly used in biochemical reactions. The resonance effect occurs because of the conduction of electrons of metal particles with photons (light particles). Many factors affect this technique including the size and shape of the metal particles as well as the materials and composition of the molecule. For the last few decades, this SPR has been a technique that can be used to analyze biomolecular interactions. In recent years, it has been proven that this surface plasmon resonance technique can be used to characterize the binding kinetics and affinities of monoclonal antibodies [8]. Some current advantages to this method include the high-quality data of epitope mapping that other technologies cannot determine at this time, most notably ELISA. SPR also provides label-free interaction analysis and is known to have the ability to study protein-carbohydrate interactions that are classified as low affinity and cannot be analyzed accurately by alternative methods [8].

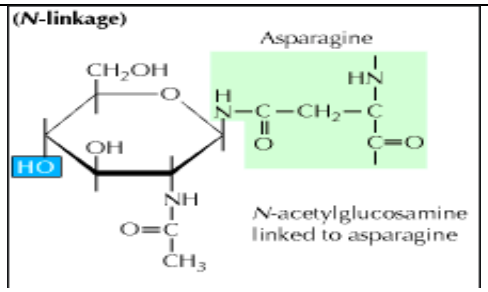
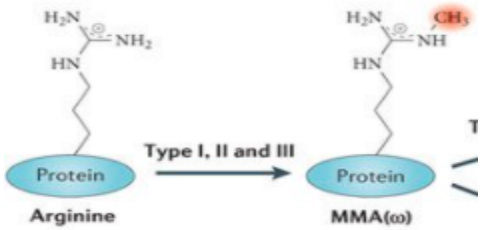
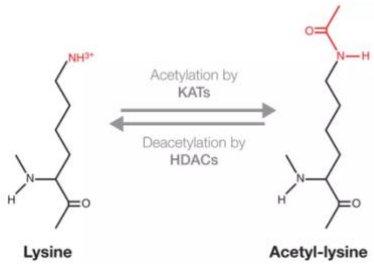
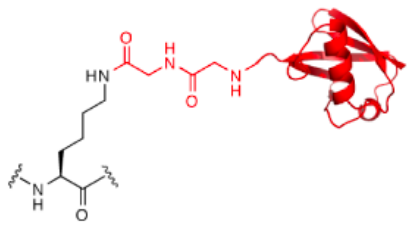
Table 1. Methodology

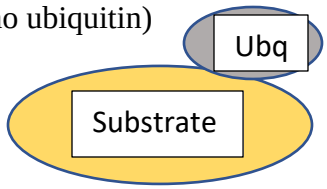
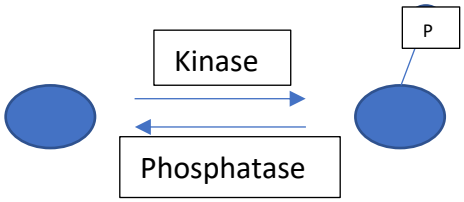
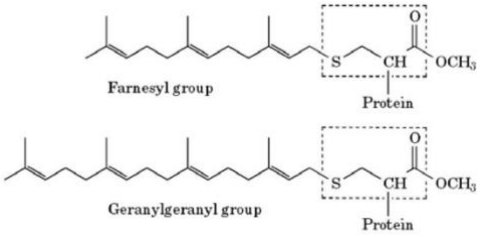
Technique	Accuracy	Method
SPR	N/A	Label-free method to analyze the kinetics of biomolecular interactions
Subnanogap sensoring	*70-90%*	A method with the ability to analyze small amino acid residues and determine their sequences
Image-Based ClpXP Digestion	*95%*	Utilizes the high frequency of Lysine (K) to determine low abundant peptides
Protein Sequencing using Mass Spectrometry	N/A	Mass-to-chare ratio of molecules to determine amino acid characteristics for comparison to a protein database for determination
Oxford Nanopore	N/A	Utilizes small nanopores that can aid in analyzing new proteins with high sensitivity and high specificity
Recognition Tunneling	>99%	Utilizes small particles to determine free amino acids

Post Translational Modifications

Proteins can go through a series of changes that can ultimately result in increased functional diversity of the proteome. These changes are referred to as post-translational modifications (PTMs). Essentially, PTMs occur by the addition of functional groups to protein sequences, deletion or degradation of proteins, or cleavage of subunits. Many different types of PTMs can occur including phosphorylation, glycosylation, ubiquitination, nitrosylation, methylation, acetylation, proteolysis, and lipidation. These PTMs are generally covalent or enzymatic modifications of proteins directly after their synthesis. These proteins utilized ribosomes (a cell organelle) to translate mRNA into polypeptide chains. These complete peptide chains may undergo PTMs to become mature proteins. These changes may be reversible or irreversible. As previously mentioned, there are 20 different amino acids that are synthesized into polypeptides however, proteins are known to contain approximately 140 different residues due to these PTMs. Ultimately, these PTMs can occur at any stage during the life of the peptide chain. These PTMs may occur fairly early on and have the potential to alter how the protein is folded. In this instance, it could affect the protein's conformational changes and stability, which could determine its functional fate. Because PTMs alter catalytic activity, it can change the proteins' biological function and activity.

Table 2. Types of Post-Translational Modifications that can occur in peptides.

PTM	Addition	Diagram	Characteristics
Glycosylation	Carbohydrate moiety to a protein		Significant effects on protein folding, conformation, stability, and activity
Methylation	Methyl group		Increases hydrophobicity of the protein and can neutralize negative charge on amino acid
Acetylation	Acetyl group		Occurs in almost all eukaryotic proteins via reversible and irreversible mechanisms
SUMOylation	Small ubiquitin-like modifier		Involved in transcriptional regulation, apoptosis, and progression through the cell cycle

Ubiquitination	Ubiquitin	(mono ubiquitin) 	Utilizes Ubiquitin polypeptide of 76 amino acids
Phosphorylation	Phosphorous		Regulates cell cycle, growth, apoptosis, and signal transduction pathways
Prenylation (lipidation)	Farnesyl or geranylgeranyl isoprenoid		Increases hydrophobicity of proteins and membrane affinity

Conclusion

After significant research, it has been noted that there is no single method to sequence proteins. Each method that is been carried out by companies and individual researchers has reached a plateau in its ability. The limitations for each method and technique are difficult ones to overcome and cannot be overlooked. From extensive research, I have concluded that the best method to seek the strongest results effectively is to combine the highlights of various currently established methods to set the standard for NGPS. Currently, the technique with the best overall outcome is the Open SPR being done by Nicoya Life Sciences. However, despite its unique flexibility to be able to work with many different types of molecules, the technology still falls short in other areas. For one, the accuracy of the results that are being produced still isn't documented, especially in comparison to other methods such as Recognition Tunneling with *greater* than 99% accuracy. Another shortcoming is that SPR has not been proven to be compatible with all materials and varying compositions. This places limitations on experimentation with different materials that could be useful in defining the complete accuracy of NGPS. SPR is also not able to sequence all sizes and shapes of molecules, again placing more limitations on researchers. Despite these current setbacks with Open SPR, it still proves to be very close to the standard we hope to set for NGPS. Moreover, one technique that could overcome these difficulties with SPR is, potentially, the work being done by Oxford Nanopore. Because the Oxford Nanopore is a company that uses technology that can deliver results with high specificity and sensitivity, it can bypass the uncertainty in SPR accuracy. Additionally, because this technique utilizes aptamers that don’t bind to all nanopores, we can easier

differentiate between target and non-target proteins. This can increase the effects of SPR simply because SPR has been used to discover new drugs. If we can use the effects of the Oxford Nanopore to identify and sequence target proteins, this could open more efficient ways to create new drugs for various diseases and illnesses once combined with Open SPR. Overall, a combination of two or more of the current techniques that have been established in recent years could create what will be known as Next Generation Protein Sequencing.

Future Work

Although NGPS has the potential to be the future of diagnostic medicine and research, there will be a great deal of time before that can happen. Because NGPS can provide such accuracy after PTMs, it makes this methodology and technology extremely sensitive. Ultimately, determining the best method for analyzing a protein sequence incorporates many different mechanisms and technologies. Surface plasmon resonance (SPR) can serve as a method to analyze protein-carbohydrate interaction that has been classified as low abundance. Through the development of SPR work, there is a possibility of it becoming the primary method for protein sequencing in conjunction with Edman degradation to make the side chains available for study. Mass spectrometry (MS) is another emerging method that provides great specificity with little room for error. Despite this fact, because the studied sequence results are compared to the proteomic

database for determination, there is the possibility for less accurate identification of amino acids. Next-generation sequencing (NGS) is bringing about new techniques for analyzing the samples and answering questions that previous methods have left researchers wondering about for years. Sanger sequencing and gel electrophoresis for DNA sequencing proved to be much less cost-effective than NGS when analyzing large samples. Through the incorporation of Next-generation DNA sequencing, researchers and scientists were able to study entire genomes of biological organisms through ultra-high throughput, scalability, and speed technology which has completely revolutionized DNA sequencing. Unfortunately, protein sequencing is far behind next-generation DNA sequencing. Currently, protein sequencing is needing to overcome issues with methodology as well as sequencing low abundance proteins. The difference in each method relates to efficiency, timeliness, and cost. When determining the best method for isolation, the reaction of the amide groups with carboxylic acids proved to be the most cost-effective mechanism. Edman degradation is what allows for these side chains to be available. Another method that is currently underway is determining which buffers are the most useful and if aqueous solutions are more effective than organic solvents.

References

1. Campbell, Molly. "Developing the World's First Single-Molecule Protein Sequencer." *Proteomics & Metabolomics from Technology Networks*, 24 Apr. 2020, www.technologynetworks.com/proteomics/blog/developing-the-worlds-first-single-molecule-protein-sequencer-333913.
2. Sponsored Content by Rapid NoVo IncJan 9 2018, Mingjie. "Deciphering Antibodies with Next-Generation Protein Sequencing Technology." *News*, 29 Oct. 2018, www.news-medical.net/news/20180109/Deciphering-Antibodies-with-Next-Generation-Protein-Sequencing-Technology.aspx.
3. "Review: Antibody Protein Sequence Analysis Using Mass Spectrometry." *Rapid Novor Inc*, 16 July 2019, www.rapidnovor.com/resources/review-antibody-protein-sequence-analysis-using-ms/.
4. Nasykhova, Yulia A, et al. "Recent Advances and Perspectives in next Generation Sequencing Application to the Genetic Research of Type 2 Diabetes." *World Journal of Diabetes*, Baishideng Publishing Group Inc, 15 July 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6656706/.
5. Cerf, Marlon E. "Beta Cell Dysfunction and Insulin Resistance." *Frontiers in Endocrinology*, Frontiers Media S.A., 27 Mar. 2013, www.ncbi.nlm.nih.gov/pmc/articles/PMC3608918/#:~:text=With%20beta%20cell%20dysfunction%2C%20insulin,1).
6. Kintzing, James R, et al. "Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment." *Trends in Pharmacological Sciences*, U.S. National Library of Medicine, Dec. 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC6238641/.
7. Qin, Dahui. "Next-Generation Sequencing and Its Clinical Application." *Cancer Biology & Medicine*, Chinese Anti-Cancer Association, Feb. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6528456/.
8. https://link.springer.com/chapter/10.1007/978-3-7091-0870-3_17
9. "Protein Analysis." *Oxford Nanopore Technologies*, 10 June 2020, nanoporetech.com/applications/protein-analysis.
10. Watson, James D. *Molecular Biology of the Gene*. Benjamin-Cummings Publishing Company, 2014.